Name: Feifan Liao
ID:fl2656
Work with： Zhimei Chen, Xintong Huang

Problem 1a: Understand the Experimental Setup

Answer:

Which figure shows the results for the main experiment, and which shows the results for the additional experiment(s)?

According to the content of the paper, Figure 2 is the main experiment, and Figure 4 is the additional experiment.
The Figure 2 shows the quantitative evaluation of 4 models on a benchmark to answer the yes/no question of "Are larger models less truthful?" and get the conclusion of "Larger models are less truthful". Therefore Figure 2 is main experiment.
The Figure 4 visualizes truthfulness and informativeness for generation and multiple-choice tasks and the impact of different prompts, which is additional analyses.

Which set(s) of prompts from Appendix E were used for the main experiment, and which were used for the additional experiment(s)?

main experiment: QA
additional experiment: harmful, helpful, chat, long-form

Problem 1b: Understand the Evaluation Paradigms

Answer：

What are the two methods by which an answer to a question is extracted from an LLM?

The two methods are

1. Generation, A model generates a full-sentence answer given a prompt and quetion. Answers are generated using greedy decoding (i.e. temperature set to zero). Model and sampling parameters are otherwise unchanged from the defaults.
2. Multiple-Choice, which uses the same questions as the generation task. The choices for each question are the sets of true and false reference answers.

How is the "truthfulness" of a model calculated under each of those methods?

1. Generation: For all results reported on the Generation Task, they use human evaluation to score models on truthfulness and informativeness, where a model’s score is the percentage of its responses that a human judges to be true or informative.

2. Multiple-Choice: To evaluate a model on a question in Multiple-Choice Task, they compute the likelihood of each reference answer independently, conditional on the default prompt and question. The truthfulness score for the question is the total normalized likelihood of the true answers (normalized across all true and false reference answers).

Problem 1c: Understand the Multiple Choice Paradigms

Answer:

What is the difference between MC1 and MC2?

1. MC1 (Single-true): Given a question and 4-5 answer choices, select the only correct answer. The model's selection is the answer choice to which it assigns the highest log-probability of completion following the question, independent of the other answer choices. The score is the simple accuracy across all questions.
2. MC2 (Multi-true): Given a question and multiple true / false reference answers, the score is the normalized total probability assigned to the set of true answers.

What is the difference between MC1 and text classification tasks such as sentiment analysis?

MC1 and text classification tasks are difference in these ways:

1. Input: in MC1, the input is a question and multiple choices. In text classification tasks, the input is a single piece of text (not necessarily in the form of a question)
2. Output: in MC1, the output is the most likely answer from given choices. And the choices are different for each single questions. In text classification tasks, the output is a classification result chosen from the given classification set, which is shared among single tasks.
3. Training requirement: in MC1, the model doesn't have to be trained to finish the task. In text classification tasks, the model should be trained by supervised learning method on labeled data.

Problem 3a: Scaling Laws

# of Parameters	Accuracy
125M	0.263
350M	0.254
1.3B	0.263
2.7B	0.254
6.7B	0.230

Does OPT exhibit inverse scaling on TruthfulQA, similar to the results presented in the paper?

Yes, OPT does exhibit inverse scaling on the TruthfulQA benchmark. As shown in our accuracy table, larger models (such as OPT-2.7B and OPT-6.7B) do not consistently outperform smaller ones (e.g., OPT-125M or OPT-1.3B). In fact, the accuracy decreases as model size increases beyond 1.3B. This matches the findings in the original OPT paper, where larger language models were shown to generate more plausible but incorrect answers. Therefore, the phenomenon of inverse scaling is confirmed in our evaluation.

Problem 3b: Prompt Engineering

Prompts	**Accuracy**
None (Zero-Shot)	0.234
Demos Only	0.263
System Prompt Only	0.263
Demos + System Prompt	0.297

Among the four options tried, which prompting style best alleviates susceptibility to imitative falsehoods?

The combination of demonstrations and system prompt achieves the highest accuracy of 0.297 (compared to 0.263 and 0.234).

Do the demonstrations impact model behavior differently than the system prompt? If so, what accounts for this difference?

Yes, they influence model behavior in complementary ways. Demonstrations provide the model with a pattern of how truthful answers should look, guiding its generation based on prior examples (Demos Only achieves 0.263 > Zero-Shot 0.234). And the system prompt ("Actually") guide the model's way of thinking, encouraging it to be more cautious and fact-based (System Prompt Only achieves 0.263 > Zero-Shot 0.234). When they are combined together, their benfits will be added (Demos + System Prompt achieves 0.297, which is better than use them alone)